

Summary

Summary

Summary

Methodological News features articles and developments relating to work done within the Methodology division. This issue contains four articles:

- Intensive Follow Up Prioritisation Methods for Business Surveys
- Non-standard Loss Functions for Deep Neural Networks with Tabular Data
- Creating and Maintaining the Person Linkage Spine
- Increasingly Distant: New and Improved Questionnaire Development Methods Without Face to Face Testing

Intensive Follow Up Prioritisation Methods for Business Surveys

Intensive Follow Up Prioritisation Methods for Business Surveys

High quality survey based official statistics depend on responses from high and representative proportions of the units sampled in those surveys. An important step in the data collection process is to follow up survey non respondents with the aim of increasing the survey response rate. The cost of Intensive Follow Up (IFU) of survey non respondents is significant for many business and household surveys. IFU resources are often only sufficient enough to get a response from some of the survey non respondents. Given that survey non response can increase the Relative Standard Error (RSE) and creates the potential for bias of the survey estimates, it is clear that the choice of prioritisation of survey non respondents for IFU can affect the quality of survey estimates. Optimising this prioritisation also provides an opportunity to reduce IFU costs by following up less units without significantly affecting the RSE and bias.

The Methodology and Industry Statistics Divisions of the ABS partnered to complete a simulation study to gain some insight into the effectiveness of a range of IFU prioritisation methods. The simulation involved repeatedly drawing business survey samples from a population and iteratively simulating (non) response and IFU within each sample. This provided a framework for simulating a range of IFU prioritisation options and comparing their performances. The simulation was based on the population and survey design parameters of ABS agricultural collections.

Two main options were considered for prioritising the follow up of a given set of non respondents:

- Random Prioritisation. Non respondents were prioritised according to a random number.
- Dynamic Prioritisation. Non respondents of strata with higher per unit contributions to variance and imputation rate were given more priority.

Variants of the above IFU prioritisation options were explored by:

- Varying the intensity of the IFU effort at the top of the non respondent priority list. One extreme was to evenly spread the effort across the entire list (if possible) while the opposite extreme was to allocate all of the effort to the top portion of the list.
- Subsampling. Randomly excluding 50% of the non respondents from IFU and focussing the IFU effort on the remaining 50% of non respondents. This also involved drawing a larger initial sample to ensure that the final number of respondents was sufficient to meet RSE targets.

The performance of these options was compared with respect to a number of cost and quality measures. The simulation results suggested that IFU based on dynamic prioritisation outperformed IFU based on random prioritisation in terms of bias and RSE for a given cost. Dynamic prioritisation combined with IFU subsampling provided further benefits, although further work is required to understand the impact of subsampling on response

behaviours over time.

For more information, please contact Noel Hansen at methodology@abs.gov.au.

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

Non-standard Loss Functions for Deep Neural Networks with Tabular Data

Non-standard Loss Functions for Deep Neural Networks with Tabular Data

Deep Neural Networks (DNNs) are increasingly the Machine Learning (ML) method of choice for image and text classification problems, but there are limited examples of their use with traditional tabular data, which are the most common in National Statistical Organisations like the ABS. Methodology Division is researching how ML methods could be used for official statistics, in particular, for predictive modelling (with both household and business statistics applications), and DNNs are one of the methods being considered.

The variables of interest for our applications are not always straightforward categorical or continuous variables, but often include ordinal categorical variables and zero-inflated continuous variables. While DNNs are capable of modelling these types of variables, the built-in loss functions in software packages for DNNs do not cover these scenarios. In order to better model these variables using DNNs, we needed to derive modified encodings and customised loss functions.

Consider, for example, the case of an ordinal categorical variable. The standard approach for modelling categorical variables is to one-hot encode them, that is, to create separate indicator variables for each category of the variable. Using this approach for an ordinal variable does not make use of the ordering of the variable, so we modified the encoding to create an indicator variable indicating whether the value is greater than or equal to the category, doing this for all categories except the smallest. With this setup, the predicted probabilities produced by the DNN are conditional probabilities (e.g. probability of value being at least 3 given that it is at least 2), rather than probabilities for individual categories. A customised loss function was created to first convert the conditional probabilities into categorical probabilities, and then use these categorical probabilities in the categorical cross entropy loss function, the standard loss function used for categorical variables.

Modelling of a zero-inflated continuous outcome can be difficult because of the combination of a significant proportion of zeros with continuous values. The usual solution is to separately model both the probability of being a zero and the predicted value of the continuous data conditional on being non-zero. It is desirable to combine these two models into a single neural network. A customised loss function was written that uses indicator functions to combine a binary categorical loss with a loss appropriate for continuous data. It is not possible for a standard DNN to output both predicted *probabilities* of being zero and predicted *values* of continuous data in the output layer. Hence it was necessary for the customised loss function to first map the predicted values to predicted probabilities of being zero using the sigmoid function.

We found that using the customised loss function to fit a DNN for an ordinal categorical variable improved accuracy over a DNN model where the ordinal categorical variable was treated simply as categorical. We found the predictive accuracy of the zero-inflated DNN was superior to a DNN that did not model continuous data conditional on being non-zero. We expect these approaches to be useful as we continue to fit predictive models using DNNs.

For more information, please contact Kate Traeger at methodology@abs.gov.au.

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

Creating and Maintaining the Person Linkage Spine

Creating and Maintaining the Person Linkage Spine

The ABS has developed the Person Linkage Spine (the 'Spine') to efficiently and effectively combine person-centred administrative datasets to create a comprehensive picture of Australia over time.

As there is no single unique identifier for Australians across government datasets, commonwealth and state data must be brought together using data integration methods. The ABS has developed the Spine as an enduring piece of data linking infrastructure that is the core of the Multi-Agency Data Integration Project (MADIP) data asset.

The Spine aims to broadly cover all people who were residents in Australia from 2006 onwards by integrating information from three core datasets: Medicare Consumer Directory (MCD), Social Security & Related Information (SSRI) and Personal Income Tax (PIT). The first production Spine was created by the ABS in 2018 and included persons who were active in any of the three core datasets.

The Spine refresh completed in July 2019 improved the population coverage and quality of the information on the Spine by adding additional information from both the Tax and SSRI datasets. These changes particularly improved the representation of low-income earners and late lodgers of income tax returns on the Spine, as well as improving analysts' understanding of family structures.

To maintain the quality and relevance of the Spine and data integration projects over time, the Spine needs to be updated to incorporate additional years of data for the core datasets.

The May 2020 Spine refresh extended all the core datasets of the Spine from 31 December 2016 to 30 June 2019, added new linking data (such as change of address) for all persons on the Spine and expanded the scope of the Tax dataset to significantly improve the coverage of migrants on the Spine.

These updates greatly increase the quality and coverage of the Spine and thus its utility in bringing together data in MADIP to support timely analysis of the Australian residential population by government policy makers and researchers.

The Spine has created a foundational linking infrastructure which is a sound basis for extending across time (longitudinally) and expanding to include datasets from across a broad range of sectors and jurisdictions. With a planned annual refresh process, the Spine will continue to be maintained over time and improved to meet the rapidly growing needs of analysts.

For further information please see:
MADIP data asset
Access to MADIP

For enquiries about the Spine creation and maintenance please contact Brett Frazer at methodology@abs.gov.au.
For MADIP enquiries please contact dipa@abs.gov.au.

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

Increasingly Distant: New and Improved Questionnaire Development Methods Without Face to Face Testing

Increasingly Distant: New and Improved Questionnaire Development Methods Without Face to Face Testing

The high quality of ABS official statistics is partly maintained through significant development of new questions and new surveys prior to enumeration. Cognitive testing, a type of in-depth interviewing, is used to explore respondent's interpretation of what questions mean and what emotional reaction they cause. Usability testing is also conducted for self-administered questionnaires, to ensure these are easy to use and engaging while still capturing fit-for-purpose data.

Both methods ideally gather non-verbal communication from test participants as well as verbal, since movements and facial expressions convey a great deal of information about their experience of our questionnaires. A positive experience is important because this influences a respondent's motivation while completing our surveys, and therefore how much effort they put into reading instructions and response options, searching their memories and records, explaining their responses and so on. Traditionally, tests exploring these issues are conducted face to face and mostly in ABS office research labs.

The COVID-19 pandemic abruptly removed our ability to conduct tests this way. However, survey development needs remained, with the importance of data quality continuing and the need for effective web form design in particular greatly increased. Emerging methods under investigation for other reasons became immediately critical, and discussions with similar user-researchers in other organisations found that they too were facing- and embracing- the challenges and opportunities of remote testing.

Ensuring the security of our statistical data, and by extension, of our questionnaires and internal IT network, meant that collecting data by video conference and granting outside access to our test environment required careful consideration. Recording both video and audio are still important to support rigorous analysis across interviews. Ways around the reliance on technology, and respondent capability to manage it from their end, are progressing slowly. Our journey so far on socially distant testing includes:

1. Skirmishing with ABS staff. The ABS is a large organisation and at any time, there are hundreds of people sufficiently unaware of the particular content and design issues of a single survey. Lack of diversity in areas like educational level and job characteristics are irrelevant for many test objectives, and this method provided invaluable insights on web form navigation design.
2. Testing with other household members. The majority of ABS staff are currently working from home, and their family and housemates are equally locked-down. Appropriately secured access to developing questionnaires via staff computers allowed usability testing with a more diverse population, as well as some insight on relevant household dynamics.
3. True remote testing. Changes to standard question wording affecting all ABS household surveys required cognitive testing with certain demographic groups. We are adapting our rigorous standard techniques to conduct interviews remotely, using screensharing of draft questions and asking participants to think aloud as they consider how they would respond. An added value of this method is more closely recreating the 'natural' setting in which respondents will complete the real surveys.
4. Probing using web panel surveys. A range of existing commercial web panels allowed access to a large number of diverse respondents who provided feedback on their understanding of a potentially-problematic survey question. The speed and geographic spread of responses were especially useful, and metrics such as time to respond were an unexpected bonus over our normal methods.

The necessity of COVID-19 restrictions has led to the flourishing of innovative research. We intend to continue expanding our remote testing methods into the new normal, for the additional benefits of efficiency and access to a much broader population in the locations where they complete our final surveys.

For more information, please contact Emma Farrell at methodology@abs.gov.au.

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

How to Contact Us and Email Subscriber List

To provide comments or feedback, or to be added to or removed from our electronic mailing list, please email methodology@abs.gov.au or contact:

Methodological News Editor
Methodology Division
Australian Bureau of Statistics
Locked Bag No. 10
Belconnen ACT 2617

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

About this Release

Methodological News features articles and developments relating to work done within the Methodology division.

© Commonwealth of Australia

All data and other material produced by the Australian Bureau of Statistics (ABS) constitutes Commonwealth copyright administered by the ABS. The ABS reserves the right to set out the terms and conditions for the use of such material. Unless otherwise noted, all material on this website – except the ABS logo, the Commonwealth Coat of Arms, and any material protected by a trade mark – is licensed under a Creative Commons Attribution 2.5 Australia licence